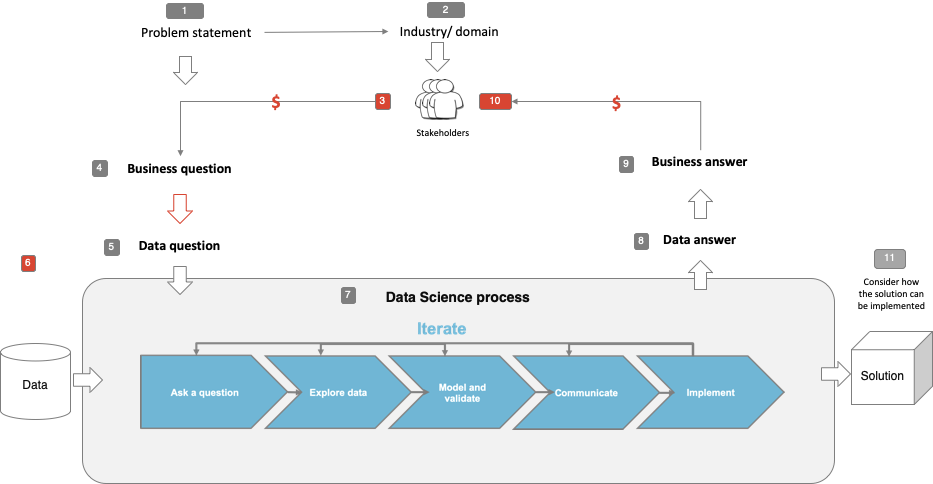
Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.



Problem statement

* What is the problem or the opportunity that the project is investigating?
  + Fake news is prevalent on the internet and people are unable to identify them as fake.
* Why is this problem valuable to address?
  + To prevent people from acting on wrong information
* What is the desired state?
  + To minimise fake news circulation

Industry/ domain

* What is the industry/ domain?
  + Social media companies
* What is the current state of this industry?
  + No regulations to prevent the spread of fake news

Stakeholders

* Who are the stakeholders? (be as specific as possible)
  + Social networking sites that allow users to share news articles (e.g. Facebook, Twitter)
* Why do they care about this problem?
  + Regulators are pressuring them to control the spread of fake news
  + Companies that advertise their products/services on these platforms could be affected by fake news if the content is against them. This might deter them from putting advertisements on these sites which would affect the social media company’s profits.
* What are the stakeholders’ expectations?
  + To be able to determine and control fake news circulation

Business question

* What is the main business question that needs to be answered?
  + Is there a way to detect fake news that is posted on the social networking sites
* What is the business value of answering this question? (quantify value and make necessary assumptions)
  + Less complications with regulators
  + Avoid losing businesses from advertisers
* What are the implications of false positives or false negatives?
  + False positive news would result in fake news circulating on social media platforms while false negative news would stop real news from circulating

Data question

* What is the data question that needs to be answered?
  + Can machine learning help to detect fake news
* What is the data required to answer the question?
  + Need real news and fake news data

Data

* Where was the data sourced?
  + Kaggle dataset

Data science process

Data analysis

* What are the highlights of the Exploratory Data Analysis (EDA)?
  + About equal number of real and fake news
  + Real news have shorter titles
  + 99% of real news come from Reuters

Modelling

* What feature engineering techniques are used?
  + Countvectorizer, TF-IDF (word, n-gram)
* What are the models used?
  + Logistic Regression, Support Vector Machine, Naive Bayes
* What are the model performance metrics?
  + Accuracy score
* Which model was selected?
  + LR with countvectorizer

Outcomes

* What are the main findings and conclusions of the data science process?
  + It is possible to use machine learning model to determine if a news is fake

Implementation

* What are the considerations for implementing the model in production?
  + How to extract the news from the social media sites
  + Should the news that is identified as fake be automatically removed from the social media site or have a warning label indicating that it could be fake, or should there be further tests to determine its authenticity

Data answer

* Was the data question answered satisfactorily?
  + Yes, based on results it shows that machine learning models can help detect fake news
* What is the confidence level in the data answer?
  + It is better to test with more data from different sources as real news dataset mainly comes from Reuters

Business answer

* Was the business question answered satisfactorily?
  + Yes, machine learning models can help detect fake news
* What is the confidence level in the business answer?
  + Good but need to test on larger dataset with data from different sources

Response to stakeholders

* What are the overall messages and recommendations to the stakeholders?
  + Machine learning models are useful to help detect fake news.

References

* Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)
  + Files can be found on github ()
    - Raw dataset: True.csv and Fake.csv
      * Downloaded from Kaggle: <https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset>
    - Data cleaning and processing: eda.ipynb
    - Cleaned dataset to be used for modelling: cleaned\_data.csv
    - Modelling: modelling.ipynb